ISSN 2707-1898 (print)
Український журнал інформаційних технологій
Ukrainian Journal of Information Technology
http://science.lpnu.ua/uk/ujit
https://doi.org/10.23939/ujit2022.01.029
Article received 06.05.2022 p.
Article accepted 19.05.20221 p.
UDK 004.[8+89]:519.7

Correspondence author
I. F. Povkhan
igor.povkhan@uzhnu.edu.ua

І. Ф. Повхан
Ужгородський національний університет, м. Ужгород, Україна

ПРОБЛЕМА ЗБІЖНОСТІ ПРОЦЕДУРИ ПОБУДОВИ КЛАСИФІКАТОРІВ У СХЕМАХ ЛОГІЧНИХ І АЛГОРИТМІЧНИХ ДЕРЕВ КЛАСИФІКАЦІЇ

Розглядається проблема збіжності процедури синтезу схем класифікаторів у методах логічних і алгоритмічних дерев класифікації. Запропонована верхня оцінка складності схеми дерева алгоритмів у задачі апроксимації масиву реальних даних набором узагальнених ознак з фіксованим критерієм зупинки процедури розгалуження на етапі побудови дерева класифікації. Даний підхід дає змогу забезпечити необхідну точність моделі, оцінити її складність, знизити кількість розгалужень та досягти необхідних показників ефективності. Вперше для методів побудови структур логічних і алгоритмічних дерев класифікації дана верхня оцінки збіжності побудови дерев класифікації. Запропонована оцінка збіжності процедури побудови класифікаторів для структур ЛДК/АДК дає можливість будувати економні та ефективні моделі класифікації заданої точності. Метод побудови алгоритмічного дерева класифікації базується на поетапній апроксимації начальної вибірки довільного об'єму та структури набором незалежних алгоритмів класифікації. Даний метод при формуванні поточної вершини алгоритмічного дерева, вузла, узагальненої ознаки забезпечує виділення найбільш ефективних, якісних автономних алгоритмів класифікації з початкового набору. Методи синтезу логічних і алгоритмічних дерев класифікації були реалізовані в бібліотеці алгоритмів програмної системи "ОРІОН ІІІ" для розв'язку різноманітних прикладних задач штучного інтелекту. Проведені практичні застосування підтвердили працездатність побудованих моделей дерев класифікації та розробленого програмного забезпечення. В роботі наведена оцінка збіжності процедури побудови схем розпізнавання для випадків логічних і алгоритмічних дерев класифікації в умовах слабкого та сильного розділення класів початкової начальної вибірки.

Ключові слова: логічне дерево; алгоритмічне дерево; класифікатор; розпізнавання образів; ознака; начальна вибірка.

Вступ / Introduction

Задачі, які об'єднуються тематикою розпізнавання образів, дуже різноманітні та виникають у сучасному світі в усіх сферах економіки та соціального контенту діяльності людини, що приводить до необхідності побудови та дослідження математичних моделей відповідних систем [1], [2], [3], [4]. Станом на тепер не існує універсального підходу до їх розв'язання, запропоновано декілька досить загальних теорій та підходів, що дають змогу вирішувати багато типів (класів) задач, але їх прикладні застосування відрізняються досить великою чутливістю до специфіки самої задачі або предметної області застосування [6]. Багато теоретичних результатів отримано для спеціальних випадків та підзадач, причому варто відзначити, що вузьким місцем вдалих реальних систем розпізнавання залишається необхідність виконання величезного об'єму обчислень та орієнтація на потужний апаратний інструментарій. Концепція дерев класифікації (дерев рішень) позбавлена значної частини наведених вище недоліків та дає можливість ефективно працювати в задачах із даними довільних шкал (де інформація задається в природній формі) [7], [8], [9], [10].

На сьогодні актуальні різні підходи до побудови систем класифікації у вигляді логічних дерев та алгоритмічних класифікації (ЛДК/АДК) [11], [12], [13], причому інтерес до методів розпізнавання, які використовують ЛДК, викликаний рядом корисних властивостей, якими вони володіють. З одного боку, складність класу функцій розпізнавання (ФР) у вигляді моделей ЛДК, при визначених умовах, не перевищує складності класу лінійних функцій розпізнавання (простішого з відомих). З іншого боку, ФР у вигляді дерев класифікації дають змогу виділити в процесі класифікації як причинно-наслідкові зв'язки (та однозначно врахувати їх у подальшому), так і фактори випадковості або невизначеності, тобто врахувати одночасно і функціональні, і стохастичні відношення між властивостями та поведінкою всієї системи, причому відомо, що процес класифікації нових, таких що до сих пір не зустрічалися об'єктів світу багатьох тварин і людей (за виключенням об'єктів, інформація про які передається генетичним шляхом (наслідковим), а також у деяких інших випадках), відбувається за так званим логічним деревом рішень [14].

Зафіксуємо, що в більшості задач прогнозування та класифікації, які використовують неструктуровані дані (наприклад набори дискретних зображень або текстові масиви), штучна нейронна мережа (підібраного типу) перевершує за продуктивністю всі інші типи алгоритмів або фреймворків дерев рішень. У протилежному разі (у випадку структурованих масивів дискретних даних великого об'єму) значною мірою перевагу мають методи та алгоритми концепції дерев рішень [15], [16], [17]. В практичній площині досить часто алгоритми та методи побудови ЛДК на виході дають структурно складні логічні дерева (в плані кількості вершин, кількості розгалужень, належності до класу нерегулярних дерев), які нерівномірно заповнені даними, мають різну кількість розгалужень. Такі складні деревоподібні структури досить складно сприймаються для зовнішнього аналізу за

рахунок великої кількості вузлів (вершин) та великої кількості покрокових розбиттів початкової начальної вибірки (НВ), як містять мінімальну кількість об'єктів (можливо навіть одиничні об'єкти в найгіршому випадку). Розглянемо в роботі принципове питання щодо методів дерев класифікації (моделей класифікації) – питання збіжності процедури побудови дерева класифікації (методів дерев класифікації), структур ЛДК/АДК.

Об'єкт дослідження – процеси синтезу логічних дерев класифікації різних типів та схем.

Предмет дослідження – методи та алгоритми побудови логічних дерев класифікації (дерев рішень).

Мета роботи – визначення верхньої оцінки збіжності процедури синтезу схем логічних і алгоритмічних дерев класифікації в задачах штучного інтелекту.

Для досягнення зазначеної мети визначено такі основні *завдання дослідження*:

- аналіз структурної складності ЛДК/АДК;
- визначення верхньої межи збіжності схем ЛДК/АДК.

Аналіз останніх досліджень та публікацій. Усі базові підходи в теорії розпізнавання мають свої переваги і недоліки та утворюють єдиний інструментарій розв'язку прикладних задач теорії штучного інтелекту. Зокрема цілісно пропрацьованим з математичної точки зору є класичний алгебраїчний підхід, розроблений Ю. І. Журавльовим [19]. Цей напрямок розвитку теорії розпізнавання зв'язаний з побудовою моделей алгоритмів класифікації та вибором у рамках моделі оптимального за якістю алгоритму розпізнавання. Центральну увагу в роботі приділено актуальній концепції дерев рішень (структур ЛДК). Зокрема з робіт [1], [3], [20]] відомо, що схема класифікації, яка задається довільним підходом, методом, алгоритмом дерева класифікації має деревоподібну логічну структуру, причому структура логічного дерева складається з вершин (ознак), які групуються за ярусами та побудовані (відібрані) на певному кроці (етапі) побудови моделі дерева класифікації [21], а головна особливість деревоподібних систем розпізнавання полягає в тому, що важливість окремих ознак (групи ознак чи їх наборів) визначається відносно функції, яка задає поділу об'єктів на класи [22].

В роботі [21] пропонується схема генерації структури дерева класифікації на основі покрокової селекції елементарних атрибутів, недоліком якої є висока залежність складності моделі від ефективності фінальної мінімізації, процедури обрізки дерева. В роботах [23], [24], [25] пропонується модульна схема побудови класифікаторів у вигляді структур дерев класифікації, яка дає змогу обійти обмеження традиційних методів дерев рішень. Робота [24] пропонує ефективну схему генерації узагальнених ознак на основі побудови наборів геометричних об'єктів. Недоліком такої схеми є обмеження щодо структури початкової навчальної вибірки та неуніверсальність в прикладному плані. Питання оцінки структурної складності моделей ЛДК на етапі мінімізації піднімаються в роботі [20].

Так, з [22] відомо, що результуюче правило класифікації, яке побудоване довільним методом або алгоритмом розгалуженого вибору ознак, має деревоподібну логічну структуру. В них на перше місце виходить питання вибору якісного критерію розгалуження. Логічне дерево складається з вершин, які групуються по ярусам і які отримані на певному кроці побудови дерева розпізнавання [25]. Тут виникає питання ефективної мінімізації структури побудованої моделі дерева класифікації. Важливою задачею, яка виникає з роботи [23], є питання синтезу дерев розпізнавання, які будуть представлятися фактично деревом алгоритмів. Важливим напрямком досліджень структур ЛДК залишаються питання стосовно генерації дерев рішень для випадку малоінформативних ознак [14] та актуальне питання теорії дерев класифікації – питання можливої побудови всіх варіантів логічних дерев, які відповідають початковій НВ та відбору мінімального за глибиною, структурною складністю (кількістю ярусів) дерева класифікації [26], [27], [28], [29].

Результати дослідження та їх обговорення / Research results and their discussion

Збіжність синтезу структури логічного та алгоритмічного дерева класифікації. Нехай задана НВ в наступному вигляді:

$$\left(x_1, f_R\left(x_1\right)\right),\ldots,\left(x_m, f_R\left(x_m\right)\right). \quad (1)$$

Зауважимо, що тут $x_i \in G$, $f_R\left(x_i\right) \in \left\{0,1,\ldots,k-1\right\}$, $(i=1,2,\ldots m)$, m – кількість об'єктів з НВ, $f_R(x_i)$ – деяка скінчено-значна функція, що задає поділу R множини G на класи (образи) $H_0, H_1,\ldots,H_{k-1}$. Відношення $f_R(x_i)=l$, $\left(l=1,2,\ldots,k-1\right)$ означає $x_i \in H_l$, $x_i=\left\{x_{i1}, x_{i2},\ldots,x_{in}\right\}$, x_{ij} – значення j-гої ознаки для об'єкта x_i, $(j=1,2,\ldots,n)$, n – кількість ознак в НВ.

Отже, НВ – це сукупність (точніше послідовність) деяких наборі, причому кожний набір – це сукупність значень деяких ознак та значень деяких функцій на цьому наборі [18]. Можна ще сказати, що сукупність значень ознак – це деяке зображення, а значення функції відносить це зображення до відповідного образу. Ставиться задача побудови конструкції ЛДК/АДК –L на основі масиву початкової НВ типу (1) та визначити значення його структурних параметрів p (тобто

$$F\left(L\left(p,x_i\right), f_R\left(x_i\right)\right) \to opt\ ).$$

Нехай на кожному кроці в процесі побудови логічного дерева (деякої моделі ЛДК) буде вибиратися тільки одна відібрана елементарна ознака з набору фіксованих ознак $\left(\varphi_1,\varphi_2,\ldots,\varphi_n\right)$. Тоді на n-му кроці процедури побудови дерева класифікації схема ЛДК буде представляти собою деякий предикат p_n (узагальнену ознаку, яка побудована з набору елементарних ознак) [23], [30], який є найефективнішою апроксимацією початкової НВ загального вигляду (1) (звичайно, що це справедливо і для випадку структури АДК).

Зокрема p_n буде представляти деяку деревоподібну схему (дерево класифікації), яке складається з n вершин, тобто в структуру предикату p_n будуть входити всього n елементарних ознак (атрибутів дискретного об'єкту НВ) з початкового набору.

Зауважимо, що послідовність предикатів $p_1, p_2,\ldots, p_j$ (узагальнених ознак) збігається до початкової НВ вигляду (1), якщо, починаючи з деякого Q, буде виконуватись умова:

$$f_{Q+m} = f_R\left(x_i\right),\ \left(i=1,2,\ldots,m\right),\ \left(m \geq 0\right).$$

Деяку елементарну ознаку, яка буде вибиратися (фіксуватися) на n-му кроці в схемі побудови моделі

ЛДК, позначимо через φ_n. Зрозуміло, що ознаці φ_n відповідає деякий фіксований шлях $r_1, r_2, \ldots$, який завершується даним атрибутом (вершиною дерева класифікації – моделі ЛДК). Наприклад, на рис. 1. зображено ЛДК, в якому вершині φ_2 (ознаці) відповідає шлях $\{0\}$, а вершині φ_5 – шлях $\{0,1\}$.

Шлях, який відповідає елементарній ознаці φ_n вказаним чином, позначимо через T_n, а через D_n позначимо множину тих пар $(x_i, f_R(x_i))$ початкової НВ загального вигляду (1), для яких об'єкти w_i належать шляху T_n. Наприклад, для структури ЛДК (рис. 1), нехай $\varphi_n = \varphi_4$, тоді шлях T_n буде мати вигляд $\{1,0\}$.

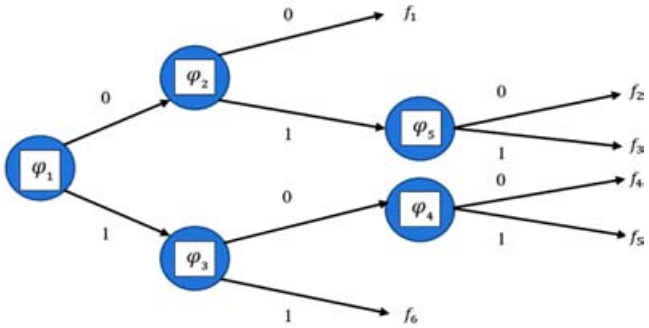

Рис. 1. Структура ЛДК з елементарними ознаками в якості вершин / LCT structure with elementary features as vertices

У такому випадку деякий об'єкт w_i належить шляху $\{1,0\}$, якщо виконуються умови: $\varphi_1(w_i) = 1$ та $\varphi_3(w_i) = 0$.

Далі будемо вважати, що елементарна ознака φ_n слабо розділяє множину D_n, якщо в D_n існують такі пари $(x_i, f_R(x_i))$ та $(x_j, f_R(x_j))$, що $\varphi_n(x_i) = 0$ та $\varphi_n(x_j) = 1$ (тобто $\varphi_n(x_i) \neq \varphi_n(x_j)$).

Кінцевою потужністю схеми методу дерева класифікації (моделей ЛДК/АДК) будемо називати кількість всіх кінцевих вершин (визначених листів) даної схеми. Наприклад, для ЛДК з рис. 1. потужність буде дорівнювати 6.

Очевидно, що кінцева потужність схеми методу дерева класифікації також дорівнює кількості всіх кінцевих шляхів у даній схемі. Зрозуміло, що індукцією за n легко довести, що кінцевою потужністю кожної з вищевказаних схем p_n (предикатів), дорівнює n+1. Дійсно, те, що кінцева потужність p_1, до складу якої входить тільки одна ознака або алгоритм (випадків ЛДК/АДК), дорівнює 2, є очевидним.

Нехай кінцева потужність схеми p_n дорівнює n+1. Підрахуємо кінцеву потужність p_{n+1}. Зрозуміло, що дана схема будується на основі схеми p_n, коли в деякій кінцевій вершині послідовно додається нова вершина (ознака, алгоритм) із номером $n+1$. Очевидно, що при додаванні цієї ознаки (алгоритму) в схему p_n зникає одна кінцева вершина та додаються дві нові кінцеві вершини. Отже, можна зробити висновок, що кількість усіх кінцевих вершин схеми p_n дорівнює $n+2$.

Припустимо, що на кожному n-му кроці процедури побудови дерева класифікації (моделі ЛДК) множина D_n слабо розділяється деякою ознакою φ_n. Далі розглянемо схему p_n. У цій схемі маємо відповідно до вищезазначеного, $n+1$ кінцевих шляхів. Завдяки тому, що D_n на кожному кроці слабо розділяється, кожний такий шлях містить хоча б одну пару початкової НВ загального вигляду (1). Окрім цього очевидно, що різні кінцеві шляхи в p_n не мають спільних пар із вибірки (1).

Отже, можна зробити висновок, що схема (предикат) p_n розділяє НВ (на основі базового критерію розгалуження введеного поточним методом дерева класифікації) на $n+1$ непустих частин (підмножин), що не перетинаються. Оскільки в початковій НВ всього знаходяться m навчальних пар, то схема p_{m-1} (або предикат з меншим номером) повністю розділить початкову НВ, тобто p_{m-1} буде повністю розпізнавати вибірку.

Отже, якщо на кожному n-му кроці відібрана елементарна ознака φ_n слабо розділяє множину D_n, то в цьому випадку процес побудови ЛДК збігається відносно початкової НВ та завершується не більше ніж за $m-1$ кроків, де m – кількість усіх навчальних пар початкової НВ.

Зауважимо, що умова слабого розділення класів початкової НВ є доволі слабою – тому вона забезпечує невисоку збіжність процедури побудови дерева класифікації, отже, важливо розглянути питання збіжності процесу при більш сильній умові. Тому будемо припускати, що маємо справу з випадком, коли НВ містить інформацію про два класи (образи) H_0 та H_1, а сама НВ має детерміновану природу. Нехай n_j – кількість навчальних пар $(x_i, f_R(x_i))$ у початковій НВ, які задовольняють співвідношення $f_R(x_i) = j$, $(j = 0,1)$, причому для спрощення та визначеності покладемо, що $n_0 \geq n_1$.

Зафіксувавши $f_R(x) \equiv 0$, буде отримано деяку узагальнену ознаку (схему) f_0, яка апроксимує (повністю або частково) початкову НВ. Очевидно, що в даному випадку (тобто в ситуації, коли ще не зроблено вибір жодної елементарної ознаки φ_n), узагальнена ознака (схема) f_0 є найкращою апроксимацією початкової НВ. Далі величину n_1 будемо називати безумовною кількістю помилок у початковій НВ.

Нехай на першому кроці побудови дерева класифікації відібрана (довільним шляхом) деяка елементарна ознака φ_1 – причому дана ознака розіб'є початкову вибірку на дві частини (підмножини) H_0 та H_1, де H_j – множина всіх пар $(x_i, f_R(x_i))$ початкової НВ, для яких виконується співвідношення $f_1(x_i) = j (j = 0,1)$.

Нехай n_m^j – множина всіх пар $(x_i, f_R(x_i))$ з вибірки H_j, $(j = 0,1)$, для яких виконується співвідношення $f_R(x_i) = m (m = 0,1)$. Ознаку φ_1 можна вважати узагальненою ознакою f_1 (схемою), яка побудована на першому кроці процесу побудови ЛДК.

Уведемо величину $\rho = \max(n_0^0, n_1^0) + \max(n_0^1, n_1^1)$, яка представляє собою кількість правильних відповідей (класифікацій), які реалізуються узагальненою ознакою f_1, а відповідно величина n_0 є кількістю правильних відповідей (класифікацій), які реалізуються узагальненою ознакою f_0.

Під кількістю правильних відповідей розуміємо кількість тих навчальних пар $(x_i, f_R(x_i))$ у початковій навчальній вибірці типу (1), для яких виконується співвідношення рівності $f_R(x_i) = f_1(x_i)$.

Оскільки $n_0^0 + n_0^1 = n_0$ та $n_1^0 + n_1^1 = n_1$, то будемо мати наступне:

$$\rho = \max\left(n_0^0, n_1^0\right) + \max\left(n_0^1, n_1^1\right) \geq n_0 . \quad (2)$$

Отже, при виборі ознаки φ_1 кількість правильних відповідей як мінімум не зменшується. Кількість помилок, які дає узагальнений алгоритм f_1, буде дорівнювати:

$$m - \rho = n_1 - (\rho - n_0) \leq n_1 . \quad (3)$$

Зауважимо, що (3) випливає з (2). Уведемо величину $\lambda_1 = \frac{n_1}{m - \rho}$ та назвемо її якістю елементарної ознаки φ_1 відносно початкової НВ, аналогічно визначається λ_n ознаки φ_n відносно початкової НВ $(n = 1,2,3,\ldots)$.

Потужністю деякої побудованої узагальненої ознаки (УО) або набору УО (для фіксованого кроку схеми АДК) будемо називати кількість навчальних пар $\left(x_i, f_R(x_i)\right)$ початкової НВ вигляду (1), які апроксимує (правильно класифікує) дана узагальнена ознака (послідовність узагальнених ознак).

Важливим для схем АДК є те, що при покроковому розбитті НВ на дві вибірки H_0 та H_1 (і так далі) частина вибірки буде повністю покриватися поточним алгоритмом класифікації (узагальненою ознакою або їх набором) – тобто будемо мати випадок сильного розділення класів масиву НВ. Отже, можна зробити припущення, що складність кінцевої схеми АДК (загальна кількість кроків побудови дерева) буде значною мірою залежати від процедури початкової оцінки та відбору набору незалежних алгоритмів класифікації a_i, їх початкових параметрів, параметрів наборів УО f_i, які вони генерують для кожного кроку схеми АДК.

Тоді для схеми АДК важливо розглянути загальну складність процедури побудови дерева класифікації за умови слабкої роздільності класів початкової НВ, при якій генерується не більше однієї УО потужністю в одиницю для кожної вершини дерева та умови сильної роздільності, коли обмежень на кількість УО та їх потужність не накладається умовами задачі та практичною доцільністю, і можливо їх будувати.

На першому етапі розглянемо випадок слабкого розділення класів з обмеженнями на набори УО, що будуються, схемою АДК. Відзначимо, що процедура побудови алгоритмічного дерева має певні особливості з точки зору поетапної апроксимації початкової НВ послідовністю УО – нехай на кожному кроці побудови деякої моделі АДК буде вибиратися для роботи один фіксований алгоритм класифікації з набору відібраних алгоритмів $(a_1, a_2, \ldots, a_n)$, причому дерево класифікації може бути побудовано одним алгоритмом a_i та послідовністю УО, які він генерує.

Отже, після проведення n кроків процедури побудови дерева класифікації структура АДК буде представляти собою деяку схему s_n (узагальнену ознаку другого порядку, яка побудована з набору синтезованих алгоритмами класифікації УО), яка є найбільш ефективною апроксимацією початкової НВ загального вигляду (1) набором незалежних алгоритмів класифікації та їх УО. Зокрема s_n буде представляти деяку деревоподібну схему (структуру ДУО), яка складається з n вершин, тобто в конструкцію схеми s_n будуть входити всього n алгоритмів класифікації (УО – при умові генерації для кожного кроку процедури побудови дерева не більше однієї узагальненої ознаки мінімальної потужності в одиницю) з початкового набору.

На наступному етапі дослідження для структури ЛДК зробимо припущення – якість λ_n елементарної ознаки φ_n відносно масиву початкової НВ не менше деякого числа y, де $y > 1$.

Проаналізуємо складність процедури побудови дерева класифікації при даній умові $(y > 1)$, для цього оцінимо кількість кроків, за яку даний процес (процедура) реалізує повне розпізнавання масиву початкової навчальної вибірки.

Розглянемо для визначеності наступну схему побудови дерева класифікації (рис. 2).

Нехай n_1 – безумовна кількість помилок початкової НВ. Елементарна ознака φ_1^1 розділяє НВ на дві вибірки: H_0 та H_1. Нехай h_0 та h_1 відповідно безумовна кількість помилок у вибірках H_0 та H_1. Ознака φ_1^2 розділить множину H_0 на дві множини H_{00} та H_{01}. Нехай h_{00} та h_{01} – безумовна кількість помилок у вибірках H_{00} та H_{01}. Аналогічно визначимо множини H_{10}, H_{11} та кількості h_{10} і h_{11} для елементарної ознаки φ_2^2.

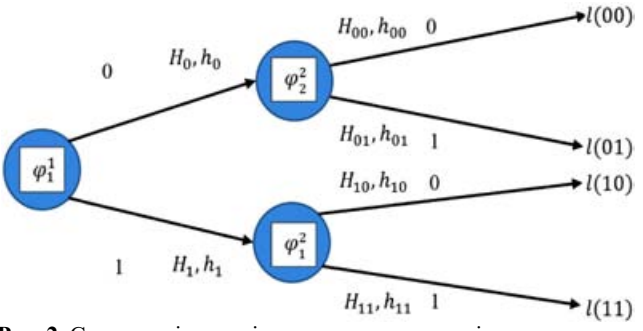

Рис. 2. Схема поділу на підмножини в структурі дерева класифікації / Scheme for splitting into subsets in the classification tree structure

З початкової умови $(y > 1)$ випливає:

$$\begin{cases} h_0 + h_1 \leq n_1 / y; \\ h_{00} + h_{01} \leq h_0 / y; \\ h_{10} + h_{11} \leq h_1 / y. \end{cases} \quad (4)$$

З виразу (4) отримаємо наступне:

$$h_{00} + h_{01} + h_{10} + h_{11} \leq n_1 / y^2 . \quad (5)$$

Зробимо такі припущення в даному відношенні: $h_0 \geq 1$, $h_1 \geq 1$, $h_{00} \geq 1$, $h_{01} \geq 1$, $h_{10} \geq 1$ та $h_{11} \geq 1$. Звідси будемо мати наступне:

$$2^1 \leq n_1 / y , \; 2^2 \leq n_1 / y^2 . \quad (6)$$

Аналогічно для набору ознак $\varphi_1^i, \varphi_2^i, \ldots$, які розташовані на i-му ярусі логічного дерева, будемо мати:

$$2^i \leq n_1 / y^i \text{ або } (2y)^i \leq n_1 . \quad (7)$$

Звідси можна зробити висновок, що процес побудови дерева класифікації буде продовжуватися до тих пір, доки в структурі дерева не буде m ярусів (рівнів), де m має наступний вигляд:

$$m = R\left(\frac{\log_2 n_1}{1 + \log_2 y}\right). \quad (8)$$

Під $R(x)$ розуміється заокруглення числа x до найближчого цілого числа, яке перевищує x. Наприклад

$$Q(1.2)=2, Q(3.7)=4, Q(4.1)=5.$$

Отже дерево класифікації, яке має m повних ярусів (тобто випадок, коли на i-му ярусі стоять 2^{i-1} вершин), має $2^{m+1}-1$ вершин – в такий спосіб розпізнавання початкової НВ при умові $(y>1)$ за допомогою повного ЛДК відбувається не більш ніж за $2^{m+1}-1$ кроків, де m розраховується за допомогою виразу (8).

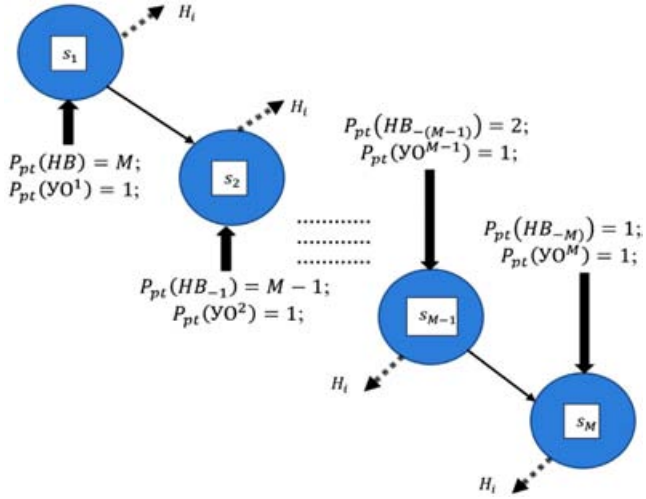

Рис. 3. Приклад структури АДК з УО в якості вершин / Example of an ACT structure with GF as vertices

Для випадку структури АДК, можна зробити висновок, що послідовність побудованих схем s_1, s_2,...,s_j (узагальнених ознак другого порядку) збігається до початкової НВ вигляду (1) не більше ніж за M кроків (де M – загальна потужність початкової НВ), навіть за умов генерації на кожному кроці тільки однієї УО, потужність кожної з яких не більше одиниці.

Деякий алгоритм класифікації, який буде вибиратися (фіксуватися) на n-му кроці в процедурі побудови моделі АДК (для генерації відповідної УО), позначимо через a_n, причому зрозуміло, що даному алгоритму a_n відповідає деяка схема s_n, яка складається з алгоритмів $a_1, a_2, \ldots, a_{n-1}$ та завершується даним атрибутом (вершиною дерева класифікації – моделі АДК). Наприклад, на рис. 3. зображено деяку модель АДК, в якій фіксованій схемі s_2 (вершині дерева класифікації, що будується) відповідає послідовність кроків (схем) $\{s_1\}$, а схемі s_M – послідовний шлях $\{s_1, s_2, \ldots, s_{M-1}\}$.

Отже, для моделі АДК можна зробити наступний висновок: схема s_n (у структурі дерева класифікації) розділяє НВ на n непустих частин (підмножин), що не перетинаються, причому оскільки в початковій НВ всього знаходиться M навчальних пар, то схема s_M повністю розділить (апроксимує) початкову НВ (тобто s_M буде повністю розпізнавати вибірку за умови генерації на кожному кроці по одній УО потужністю один). Отже, якщо на кожному n-му кроці схеми побудови АДК згенерована УО (відібраним алгоритмом класифікації a_n) слабо розділяє множину початкової НВ, то в цьому випадку процес побудови дерева класифікації збігається відносно початкової НВ та завершується не більше ніж за M кроків, де M – кількість усіх навчальних пар початкової НВ. На наступному етапі дослідження важливо розглянути випадок сильного розділення класів початкової НВ, коли жодних обмежень на алгоритми a_i щодо генерації УО не накладається (потужність побудованої УО обмежена тільки практичною можливістю самого алгоритму класифікації a_i та структурними параметрами НВ).

Нехай через $P(f_j)$ позначимо загальну потужність (апроксимаційну здатність) відповідної УО f_j, $(1 \le j \le s)$, де s – кількість УО у схемі АДК, що будується. Далі на деякому кроці $r(1 \le r \le M)$ схеми АДК побудовано послідовність узагальнених ознак $f_1, \ldots, f_r$ з відповідними їм величинами $P(f_i)=z_i$, де $(1 \le z \le M), (1 \le i \le r)$, M – загальна потужність НВ, причому серед них є величини z^{max} та z^{min}, які є для них відповідно максимальними та мінімальними (відносно поточного кроку схеми АДК). Тоді в такому випадку схему (модель) АДК буде побудовано за t кроків, де величина t визначається співвідношенням (9).

$$t \le 2 \cdot \frac{P_{pt}(HB)}{z^{max}+z^{min}} = \frac{2M}{z^{max}+z^{min}}. \quad (9)$$

Зауважимо, що у разі, коли умовою прикладної задачі на схему АДК, що будується, накладаються обмеження щодо потужності синтезованих УО (не перевищення відповідної величини P) – схему дерева класифікації (модель АДК) буде побудовано за t кроків, де величина t визначається співвідношенням

$$t \le M / P. \quad (10)$$

При жорстких обмеженнях схеми АДК на одну генеровану УО (де за умовою $P(f_j)=1, (1 \le i \le t)$), тобто у випадку слабкого розділення класів поточної задачі, схему дерева класифікації (модель АДК) буде побудовано за t кроків, де величина $t \le M$.

Табл. 1. Початкові параметри задач класифікації / Initial parameters of classification problems

Тип задачі класифікації	Розмірність ознакового простору N	Потужність масиву даних початкової НВ – M	Загальна кількість класів за поділом даних НВ – l	Відношення об'єктів різних класів НВ – (H_1/ H_2.../ H_i)
Задача класифікації геологічних даних (Z_1)	22	1250	2	756/494
Задача хімічного аналізу якості вуглеводного палива (Z_2)	14	4863	6	823/648/1412/918/583/764
Задача класифікації паводкових ситуацій басейну річки Тиса Закарпатської області (Z_3)	18	6118	3	76/108/5934
Задача класифікації паводкових ситуацій басейну річки Уж Закарпатської області (пост спостереження № 1) (Z_4)	18	4252	3	73/102/4107
Задача класифікації паводкових ситуацій басейну річки Уж Закарпатської області (пост спостереження № 2) (Z_5)	18	4139	3	68/97/3974

На наступному етапі для спрощення візьмемо задачі класифікації для яких будувалися набори структур ЛДК/АДК з робіт [19], [20], [21], [22], [23], [24], [30]. Початкові параметри даних прикладних задач представлені в наступній таблиці (табл. 1). Так в НВ представлена інформації про поділу на два класи. На етапі екзамену побудована система класифікації має забезпечити ефективне розпізнавання об'єктів невідомої класифікації відносно цих двох класів. Зауважимо, що на початковому етапі навчальна та тестова вибірка була автоматично перевірена на коректність (пошук та видалення однакових об'єктів різної належності – помили першого та другого роду).

Обговорення результатів дослідження. В (табл. 2) представлена оцінка побудованих структур дерев класифікації (ЛДК/АДК) задач з (табл. 1). Для оцінки якості побудованих класифікаторів (схем класифікації) використовувався інтегральний показник якості дерева класифікації Q_{Main} з роботи [30]. Збіжність процедури синтезу структури ЛДК/АДК оцінюється на основі кількісних показників – загальної кількості ітерацій S_{Main} та кількості ярусів структури дерева класифікації L_{Kol}. Інтегральний показник якості дерева класифікації Q_{Main} відображає базові параметри (характеристики) дерев класифікації та може бути застосована в якості критерію оптимальності в процедурі оцінки довільної деревоподібної схеми розпізнавання [20]. Відзначимо, що головна ідея методів дерев класифікації на основі автономних алгоритмів у своїй структурі, полягає в поетапній апроксимації відібраним набором алгоритмів масиву даних початкової НВ [22]. Отримані структури дерев класифікації (моделі АДК/ЛДК) з одного боку характеризується високою універсальністю відносно прикладних задач та відносно компактною структурою самої моделі, але з іншого боку вимагає істотних апаратних витрат для зберігання узагальнених ознак та початкової оцінки якості зафіксованих алгоритмів класифікації за даними НВ порівняно з нейромережевою концепцією [31], [32], [33], [34]. Тому, порівняно з концепцією АДК, методи ЛДК мають високу швидкодію схем класифікації, відносно незначні апаратні витрати для зберігання та роботи самої структури дерева та високу якість класифікації дискретних об'єктів.

Табл. 2. Порівняльна таблиця схем класифікації структур ЛДК/АДК /
Comparative table of LCT/ACT structure classification schemes

№ задачі	Метод синтезу структури дерева класифікації	Інтегральний показник якості моделі Q_{Main}	Збіжність структури дерева класифікації (кількість ітерацій) S_{Main}	Кількість ярусів структури ЛДК/АДК L_{Kol}
Z_1	Метод повного ЛДК на основі селекції елементарних ознак	0,004789	79	22
Z_1	Модель ЛДК з одноразовою оцінкою важливості ознак	0,002263	102	16
Z_2	Обмежений метод побудови ЛДК	0,003244	91	17
Z_2	Метод алгоритмічного дерева (типу I)	0,005119	46	9
Z_3	Метод алгоритмічного дерева (типу II)	0,002941	72	15
Z_3	Метод розгалуженого вибору ознак (покрокова оцінка)	0,003612	84	13
Z_4	Дерево алгоритмів (типу I)	0,005054	43	10
Z_5	Дерево алгоритмів (типу II)	0,002813	75	16

З табл. 2 можна бачити, що методи дерев алгоритмів (двох типів) показали високу швидкість збіжності процедури побудови структури дерева класифікації на представлених НВ порівняно зі схемами ЛДК. Також варто відзначити, що перший тип структури АДК показує хороший результат плані структурної складності (кількості ярусів, вершин, узагальнених ознак) побудованої моделі класифікації порівняно з логічними деревами класифікації та деревом алгоритмів другого типу. Вцілому можна зробити висновок щодо швидкої збіжності структур АДК порівняно з моделями ЛДК та перевагою за рахунок цього в структурній складності побудованого дерева класифікації та інформаційній ємності наборів узагальнених ознак.

Отже, за результатами виконаної роботи можна сформулювати такі наукову новизну та практичну значущість результатів дослідження.

Наукова новизна отриманих результатів дослідження – вперше для методів побудови структур ЛДК/АДК дана верхня оцінки збіжності побудови дерев класифікації.

Практична значущість результатів дослідження – запропонована оцінка збіжності процедури побудови класифікаторів для структур ЛДК/АДК дає можливість будувати економні та ефективні моделі класифікації заданої точності (даний метод був реалізований в бібліотеці алгоритмів програмної системи "ОРІОН ІІІ" для розв'язку різноманітних прикладних задач класифікації). Причому проведені практичні застосування підтвердили працездатність побудованих моделей дерев класифікації та розробленого програмного забезпечення.

Висновки / Conclusions

Отже, зважаючи на все вищезазначене в роботі, можна зафіксувати наступні пункти:

Для умови слабкого розділення класів у випадку ЛДК, якщо на кожному n-му кроці відібрана елементарна ознака φ_n слабо розділяє множину (підмножину) об'єктів початкової НВ, то в цьому випадку процес побудови дерева класифікації збігається відносно початкової НВ та завершується не більше ніж за m–1 кроків, де m – кількість усіх навчальних пар початкової НВ.

Дерево класифікації (структури ЛДК) за умови сильного розділення класів множини об'єктів початкової НВ, яке має m повних ярусів, рівнів (тобто випадок, коли на i-гому ярусі стоять 2^{i+1} вершин), має 2^{m+1}–1 вершин – отже розпізнавання масиву початкової НВ при умові (y>1) за допомогою повного ЛДК відбувається не більше ніж за 2^{m+1}–1 кроків, де m розраховується за допомогою виразу $m = R\left(\frac{\log_2 n_1}{1+\log_2 y}\right)$.

Загальна кількість всіх кінцевих вершин логічної структури (листів дерева розпізнавання) побудованої схеми класифікації буде однозначно визначати кінцеву

потужність схеми методу дерева класифікації (моделей ЛДК/АДК).

Потужністю деякої УО (набору побудованих УО) для фіксованого кроку схеми методу АДК вважається загальна кількість навчальних пар $(x_i, f_R(x_i))$ початкової НВ (підмножини початкової НВ) вигляду (1), які апроксимує (правильно класифікує) дана узагальнена ознака (послідовність узагальнених ознак).

У випадку слабкого розділення класів початкової НВ для схеми АДК процес побудови дерева класифікації збігається відносно масиву даних НВ та завершується не більше ніж за M кроків, де M – кількість усіх навчальних пар початкової НВ.

У випадку сильного розділення класів початкової НВ для схеми АДК, коли потужність побудованої УО (або набору УО) обмежена тільки практичною можливістю самого алгоритму класифікації a_i та початковими параметрами НВ, схему (модель) АДК буде побудовано за t кроків, де величину t можна визначити за співвідношенням (9).

References

[1] Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning*. Berlin, Springer. https://doi.org/10.1007/978-0-387-84858-7

[2] Quinlan, J. R. (1986). Induction of Decision Trees, *Machine Learning*, *1*, 81–106. https://doi.org/10.1007/BF00116251

[3] Breiman, L. L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Boca Raton, Chapman and Hall/CRC.

[4] Lupei, M., Mitsa, A., Repariuk, V., & Sharkan, V. (2020). Identification of authorship of Ukrainian-language texts of journalistic style using neural networks. *Eastern-European Journal of Enterprise Technologies*, *1-2*(103), 30–36. https://doi.org/10.15587/1729-4061.2020.195041

[5] Subbotin, S. A., & Oliinyk, A. A. (2017). The Dimensionality Reduction Methods Based on Computational Intelligence in Problems of Object Classification and Diagnosis. Szewczyk, R., Kaliczyńska, M. (eds) Recent Advances in Systems, Control and Information Technology. SCIT 2016. *Advances in Intelligent Systems and Computing*, vol 543, 11–19. Springer, Cham. https://doi.org/10.1007/978-3-319-48923-0_2

[6] Miyakawa, M. (1989). Criteria for selecting a variable in the construction of efficient decision trees, *IEEE Transactions on Computers, 38*(1), 130–141. https://doi.org/10.1109/12.8736

[7] Koskimaki, H., Juutilainen, I., Laurinen, P., & Roning, J. Two-level clustering approach to training data instance selection: a case study for the steel industry, *Neural Networks*: International Joint Conference (IJCNN-2008), Hong Kong, 1–8 June 2008: proceedings. Los Alamitos, IEEE, 2008, 3044–3049. https://doi.org/10.1109/IJCNN.2008.4634228

[8] Subbotin, S. (2013). The neuro-fuzzy network synthesis and simplification on precedents in problems of diagnosis and pattern recognition, *Optical Memory and Neural Networks, 22*(2), 97–103. https://doi.org/10.3103/S1060992X13020082

[9] Subbotin, S. A. (2013). Methods of sampling based on exhaustive and evolutionary search, *Automatic Control and Computer Sciences, 47*(3), 113–121. https://doi.org/10.3103/S0146411613030073

[10] De Mántaras, R. L. (1991). A distance-based attribute selection measure for decision tree induction, *Machine learning*, *6*(1), 81–92. https://doi.org/10.1023/A:1022694001379

[11] Karimi, K., & Hamilton, H.J. (2011). Generation and Interpretation of Temporal Decision Rules, *International Journal of Computer Information Systems and Industrial Management Applications*, *3*, 314–323.

[12] Kamiński, B., Jakubczyk, M., & Szufel, P. (2017). A framework for sensitivity analysis of decision trees, *Central European Journal of Operations Research*, *26*(1), 135–159. https://doi.org/10.1007/s10100-017-0479-6

[13] Deng, H., Runger, G., & Tuv, E. (2011). Bias of importance measures for multi-valued attributes and solutions, *Proceedings of the 21st International Conference on Artificial Neural Networks* (*ICANN*), 293–300. https://doi.org/10.1007/978-3-642-21738-8_38

[14] Subbotin, S. A. (2019). Construction of decision trees for the case of low-information features, *Radio Electronics, Computer Science, Control*, 1, 121–130. https://doi.org/10.15588/1607-3274-2019-1-12

[15] Deng, H., Runger, G., & Tuv, E. (2011). Bias of importance measures for multi-valued attributes and solutions, *21st International Conference on Artificial Neural Networks* (*ICANN*), Espoo, 14–17 June 2011: proceedings. Berlin, Springer-Verlag, *2*, 293–300. https://doi.org/10.1007/978-3-642-21738-8_38

[16] Painsky, A., & Rosset, S. (2017). Cross-validated variable selection in tree-based methods improves predictive performance, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(11), 2142–2153. https://doi.org/10.1109/TPAMI.2016.2636831

[17] Subbotin, S. A. (2014). Methods and characteristics of locality preserving transformations in the problems of computational intelligence, *Radio Electronics, Computer Science, Control*, 1, 120–128. https://doi.org/10.15588/1607-3274-2014-1-17

[18] Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques, *Informatica*, 31, 249–268.

[19] Zhuravlev, Yu. I., & Nikiforov, V. V. (1971). Recognition algorithms based on the calculation of estimates, *Cybernetics,* 3, 1–11.

[20] Vasilenko, Y. A., Vasilenko, E. Y., & Povkhan, I. F. (2003). Branched feature selection method in mathematical modeling of multi-level image recognition systems, *Artificial Intelligence*, 7, 246−249.

[21] Povkhan, I. (2020). A constrained method of constructing the logic classification trees on the basis of elementary attribute selection, *CEUR Workshop Proceedings*: Proceedings of the Second International Workshop on Computer Modeling and Intelligent Systems (CMIS-2020), Zaporizhzhia, Ukraine, April 15–19, 2020. Zaporizhzhia, 2608, 843–857. https://doi.org/10.32782/cmis/2608-63

[22] Vasilenko, Y. A., Vasilenko, E. Y., & Povkhan, I. F. (2004). Conceptual basis of image recognition systems based on the branched feature selection method, *European Journal of Enterprise Technologies*, 7(1), 13–15.

[23] Povkhan, I., & Lupei, M. (2020). The algorithmic classification trees. *Proceedings of the "2020 IEEE Third International Conference on Data Stream Mining & Processing* (*DSMP*)", August 21–25, Lviv, Ukraine, 37–44. https://doi.org/10.1109/DSMP47368.2020.9204198

[24] Povkhan, I., Lupei, M., Kliap, M., & Laver, V. (2020). The issue of efficient generation of generalized features in algorithmic classification tree methods. *International Conference on Data Stream Mining and Processing*: DSMP Data Stream Mining & Processing, Springer, Cham, 98–113. https://doi.org/10.1007/978-3-030-61656-4_6

[25] Povkhan, I. (2020). Classification models of flood-related events based on algorithmic trees. *Eastern-European Journal of Enterprise Technologies,* 6(4), 58–68. https://doi.org/10.15587/1729-4061.2020.219525

[26] Rabcan, J., Levashenko, V., Zaitseva, E., Kvassay, M., & Subbotin, S. (2019). Application of Fuzzy Decision Tree for Signal Classification. *IEEE Transactions on Industrial Informatics, 15*(10), 5425–5434. https://doi.org/10.1109/TII.2019.2904845

[27] Utgoff, P. E. (1989). Incremental induction of decision trees. *Machine learning, 4*(2), 161–186. https://doi.org/10.1023/A:1022699900025

[28] Hyafil, L., & Rivest, R. L. (1976). Constructing optimal binary decision trees is npcomplete. *Information Processing Letters, 5*(1), 15–17. https://doi.org/10.1016/0020-0190(76)90095-8
[29] Wang, H., & Hong, M. (2019). Online ad effectiveness evaluation with a two-stage method using a Gaussian filter and decision tree approach. *Electronic Commerce Research and Applications*. *35*, Article 100852. https://doi.org/10.1016/j.elerap.2019.100852
[30] Kaftannikov, I. L., & Parasich, A. V. (2015). Decision Trees Features of Application in Classification Problems. *Bulletin of the South Ural State University. Ser. Computer Technologies, Automatic Control, Radio Electronics, 15*(3), 26–32. https://doi.org/10.14529/ctcr150304
[31] Povhan, I. F. (2020). Logical recognition tree construction on the basis a step-to-step elementary attribute selection. *Radio Electronics, Computer Science, Control*, 2, 95–106. https://doi.org/10.15588/1607-3274-2020-2-10
[32] Bodyanskiy, Y., Vynokurova, O., Setlak, G., & Pliss, I. (2015). Hybrid neuro-neo-fuzzy system and its adaptive learning algorithm, *Xth Scien. and Tech. Conf. "Computer Sciences and Information Technologies"* (*CSIT*), 111–114. https://doi.org/10.1109/STC-CSIT.2015.7325445
[33] Srikant, R., Agrawal, R. (1997). Mining generalized association rules, *Future Generation Computer Systems*, *13*(2), 161–180. https://doi.org/10.1016/S0167-739X(97)00019-8
[34] Vasilenko, Y. A., & Vashuk, F. G. (2012). General estimation of minimization of tree logical structures, *European Journal of Enterprise Technologies*, 1/4(55), 29–33.
[35] Kushneryk, P., Kondratenko, Y., & Sidenko, I. (2019). Intelligent dialogue system based on deep learning technology. 15th International Conference on ICT in Education, Research, and Industrial Applications: PhD Symposium (ICTERI 2019: PhD Symposium), Kherson, Ukraine, 2403, 53–62.
[36] Kotsovsky, V., Geche, F., & Batyuk, A. (2018). Finite generalization of the offline spectral learning. *Proceedings of the 2018 IEEE Second International Conference on Data Stream Mining & Processing* (*DSMP*), Lviv, Ukraine August 21–25, 356–360. https://doi.org/10.1109/DSMP.2018.8478584

I. F. Povkhan

Uzhhorod National University, Uzhhorod, Ukraine

CONVERGENCE PROBLEM SCHEMES FOR CONSTRUCTING STRUCTURES OF LOGICAL AND ALGORITHMIC CLASSIFICATION TREES

The problem of convergence of the procedure for synthesizing classifier schemes in the methods of logical and algorithmic classification trees is considered. An upper estimate of the complexity of the algorithm tree scheme is proposed in the problem of approximating an array of real data with a set of generalized features with a fixed criterion for stopping the branching procedure at the stage of constructing a classification tree. This approach allows you to ensure the necessary accuracy of the model, assess its complexity, reduce the number of branches and achieve the necessary performance indicators. For the first time, methods for constructing structures of logical and algorithmic classification trees are given an upper estimate of the convergence of constructing classification trees. The proposed convergence estimate of the procedure for constructing classifiers for LCT/ACT structures makes it possible to build economical and efficient classification models of a given accuracy. The method of constructing an algorithmic classification tree is based on a step-by-step approximation of an initial sample of arbitrary volume and structure by a set of independent classification algorithms. When forming the current vertex of an algorithmic tree, node, or generalized feature, this method highlights the most efficient, high-quality autonomous classification algorithms from the initial set. This approach to constructing the resulting classification tree can significantly reduce the size and complexity of the tree, the total number of branches, vertices, and tiers of the structure, improve the quality of its subsequent analysis, interpretability, and ability to decompose. Methods for synthesizing logical and algorithmic classification trees were implemented in the library of algorithms of the "Orion III" software system for solving various applied problems of artificial intelligence. Practical applications have confirmed the operability of the constructed classification tree models and the developed software. The paper estimates the convergence of the procedure for constructing recognition schemes for cases of logical and algorithmic classification trees under conditions of weak and strong class separation of the initial sample. Prospects for further research and testing may consist in evaluating the convergence of the ACT synthesis procedure in a limited method of the algorithmic classification tree, which consists in maintaining a criterion for stopping the procedure for constructing a tree model by the depth of the structure, optimizing its software implementations, introducing new types of algorithmic trees, as well as experimental studies of this method for a wider range of practical problems.

Keywords*:* logical tree; algorithmic tree; classifier; pattern recognition; attribute; training sample.

Інформація про автора:

Повхан Ігор Федорович, д-р техн. наук, доцент, кафедра програмного забезпечення. **Email:** povkhan.igor@uzhnu.edu.ua; https://orcid.org/0000-0002-1681-3466

Цитування за ДСТУ: Повхан І. Ф. Проблема збіжності процедури побудови класифікаторів у схемах логічних і алгоритмічних дерев класифікації. *Український журнал інформаційних технологій.* 2022, т. 4, № 1. С. 29–36.

Citation APA: Povkhan, I. F. (2022). Convergence problem schemes for constructing structures of logical and algorithmic classification trees. *Ukrainian Journal of Information Technology, 4*(1), 29–36. https://doi.org/10.23939/ujit2022.01.029